Statistical Challenges of Cosmic Microwave Background Analysis

Benjamin D. Wandelt*
University of Illinois at Urbana-Champaign, IL 61801, USA

The Cosmic Microwave Background (CMB) is an abundant source of cosmological information. However, this information is encoded in non-trivial ways in a signal that is difficult to observe. The resulting challenges in extracting this information from CMB data sets have created a new frontier. In this talk I will discuss the challenges of CMB data analysis. I review what cosmological information is contained in the CMB data and the problem of extracting it. CMB analyses can be divided into two types: "canonical" parameter extraction which seeks to obtain the best possible estimates of cosmological parameters within a pre-defined theory space and "hypothesis testing" which seeks to test the assumption on which the canonical tests rest. Both of these activities are fundamentally important. In addition to mining the CMB for cosmological information cosmologists would like to strengthen the analysis with data from other cosmologically interesting observations as well as physical constraints. This gives an opportunity 1) to test the results from these separate probes for concordance and 2) if concordance is established to sharpen the constraints on theory space by combining the information from these separate sources.

1. Overview

What is cosmic microwave background (CMB) statistics and what is challenging about it?[1] It involves estimating the covariance structure of of a spatial random field with 10^6–10^8 pixels, given only ONE realization of this field. The covariance matrix of these pixels is a complicated non-linear function of the physical parameters of interest. Of these physical parameters there are between 10 and 20, so even finding the maximum likelihood point is hard—determining and summarizing confidence intervals around the maximum likelihood point is very non-trivial. Cosmologists want to do all this *and* have the option of building in exact or approximate physical constraints on relationships between parameters. In addition, since collecting cosmological data is so difficult and expensive we want to combine all available data sets—both to test them for mutual disagreement which might signal new physics, and to improve the parameter inferences. In all of this the quantification of the uncertainties in the results is extremely important—after all the stated significance of our results will either drive or stop theoretical investigations and the design of new observational campaigns.

Before I get on to CMB specifics in section 2 let me give you the short version of (most) of this talk for statisticians: "The CMB is an isotropic (homoschedastic) Gaussian random field s on the sphere. The desired set of cosmological parameters $\Theta = \{\theta_i\, i = 1, ..., n\}$, are related in a non-linear way to the spatial covariance structure $S_{ij} \equiv \langle s_i s_j \rangle$ of the field. Observers present us with a sampled, noisy, filtered and censored/polluted measurement of this field in several 'colors'. The analysis task is two-fold: infer the covariance structure of the field s. Infer the parameters Θ." This is what could be termed "canonical" CMB analysis.

In this talk I will mainly describe challenges presented by this canonical CMB analysis. After a brief review of the scientific motivation for studying the CMB in section 2 I will describe the form of CMB data as well as the current status and prospects of obtaining it in section 3. Section 4 then outlines a framework for extracting cosmologically useful information from the data and section 5 illuminates some examples of challenges that arise when implementing this framework. I will touch on statistical questions concerning "non-canonical" CMB analysis in section 6 and then conclude in section 7.

So why are we interested in facing the statistical challenges of CMB analysis?

2. What can we learn from the CMB?

Cosmologists are interested in studying the origins of the physical Universe. In order to do so they have to rely on data. For cosmologists, one of the great practical advantages of Einstein's relativity over Newtonian physics is the fact that we *cannot help but look into the past*. Therefore, by observing light that reaches us from farther and farther away, we can study the Universe directly at earlier and earlier times, at least to the extent to which the Universe is transparent to light. Since the early Universe was a hot and opaque plasma we can only see back to the time when the plasma cooled sufficiently (due to the Hubble expansion) to combine into neutral atoms and the mean free time between photons collisions became of order of the present age of the Universe. Photons that we observe today which scattered for the last time in the primordial plasma *are* the CMB.

*NCSA Faculty Fellow

[1] For online material relating to this talk please refer to www-conf.slac.stanford.edu/phystat2003/talks/wandelt/invited/

The change from plasma to gas, happened when the Universe was approximately 380,000 years old. CMB photons emitted at this time are therefore the most direct messengers that we can detect today of the conditions present in the Universe shortly after the Big Bang.[2] They constitute a pristine snapshot of the infant Universe which provide us with direct cosmological information uncluttered by the complex non-linear physics which led to the formation of stars and galaxies. One of the main attractions of the CMB is the conceptual simplicity with which it can be linked to the global properties of the Universe and the physics which shaped it at or near the Planck scale.

As a simple example, the serendipitous discovery of the CMB by Penzias and Wilson in 1965 showed that the CMB is isotropic to a very high degree. This was one of the key motivations for the development of the inflationary paradigm [4, 5, 6]. Inflation describes generically the emergence (from the era of quantum gravity) of a large homogeneous and isotropic Universe. Inflation also predicted the spectrum of small metric perturbations from which later structure developed through the gravitational instability (though it was not the only mechanism to do so). The corresponding anisotropies in the CMB were first convincingly detected by the DMR instrument on the COBE satellite in 1992. The rejection of alternative mechanisms for the generation of the primordial spectrum of metric perturbations in favor of inflation was a major advance driven by the measurement of the large angle CMB anisotropy. These developments have led to inflation becoming part of the current cosmological standard model. A very robust prediction of generic model implementations of inflation is the Gaussianity and homogeneity of the resulting perturbations.

Quantitatively, the properties of our Universe are encoded in a set of $n \sim 10-20$ *cosmological parameters* Θ where n depends on the level of detail of the modeling or, commonly, on the specification of theoretical priors which fix some of these parameters to "reasonable" values. These parameters specify the geometry and average energy density of the Universe, as well as the relative amounts of energy density contributed by the ingredients of the primordial soup (dark matter, ordinary baryons, neutrinos and dark energy and photons). In addition, the anisotropy carries information about the spectrum of primordial (inflationary) perturbations as well as their type (adiabatic or isocurvature). By combining observations of the anisotropy of both the effective temperature and the polarization of the CMB photons we can infer how transparent the Universe really was for the CMB photons on their way from last scattering to hitting our detectors. This in turn can tell us about the history of star formation.

A very exciting prospect is that by studying the details of CMB polarization we can infer the presence or absence of gravitational waves at the time of last scattering. A detection would offer an indirect view of one of the elusive messengers that started their journey at an even earlier epoch, adding a nearly independent constraint on the properties of the Universe at the Planck scale.

All this information is not encoded in actual features in the CMB map of (temperature or polarization) anisotropies. In fact in a globally isotropic universe the absolute placement of individual hot and cold spots is devoid of useful information. Information can, however, be stored in the invariants of the photon brightness fluctuations under the group of rotations SO(3). These are the properties of the field that only depend on the *relative* angular distance between two points of the field. For a Gaussian field, where 2-point statistics specify all higher order moments, this means that the angular power spectrum coefficients of the anisotropies contain all of the information.

The challenge for theoreticians was then to develop a detailed theory of the angular power spectrum C_ℓ, as a function of angular wavenumber ℓ, given the cosmological parameters Θ. While conceptually simple, it required a decade-long intellectual effort to model the relevant physical processes at the required level of precision. As a result, there now exist several Boltzmann codes (e.g. CMBFAST [2] or CAMB [3]), which numerically compute $C_\ell(\Theta)$ to 1% precision or better. The power spectra $C_\ell(\Theta)$ are sensitive functions of certain combinations of the parameters and weak functions of others (degeneracies). These weakly constrained parameter combinations are referred to as *degeneracies*. Within the context of the standard cosmological model, the theory of this dependence is well-understood.

It is clear from the preceding discussion that the CMB is an extremely valuable source of what amounts to "cosmological gold": information about the physics and the global properties of the early Universe. So what are the observational prospects?

3. Data from CMB Observations

A major international effort is underway to make high quality observations of the microwave sky using ground-based, balloon borne and space missions [19]. Space missions have the advantage of being able to scan the whole microwave sky. NASA's "Wilkinson Microwave Anisotropy Probe" (WMAP) was launched in 2001 with great success and is currently in operation. It reported its first year of data earlier this year [16]. WMAP will continue to collect data for at least

[2] There are other messengers, namely neutrinos and gravitational waves, reaching us from even earlier times but we do not (yet) have the technology to detect them in relevant quantities.

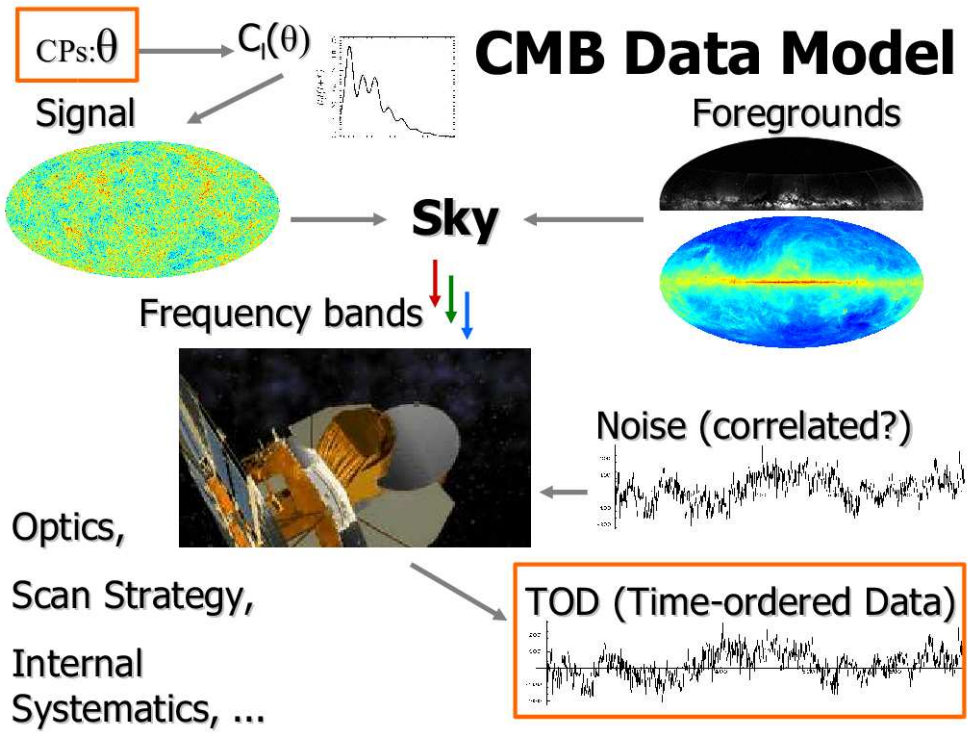

Figure 1: A schematic of how the cosmological parameters Θ (top left) are linked to the time ordered data CMB experiments actually observe (bottom right). Please see the discussion in the text.

another three years. In the medium term (ie. in late 2007) we anticipate the launch of "Planck," a joint ESA/NASA space mission.[3] Additional space missions may follow that will focus on measuring the polarization anisotropies in the CMB. In the meantime ground and balloon-based missions are jostling to accelerate our learning curve by providing maps at high angular resolution on small patches of the sky (up to a few degrees large). As a result, by the end of this decade we will have a mountain of CMB data.

From this mountain (and of course from the part of it which we have already available today) we would like to extract the cosmological gold. In order to do so we need to understand how the data and the information are related. In Figure 1 I show a simplified schematic of this relationship and I will now go through the various steps.

Starting with a set of cosmological parameters Θ we can use a Boltzmann code to compute the power spectrum C_ℓ. Given this power spectrum and the assumption of Gaussianity and isotropy we have all the information we need to create a statistical realization of a CMB anisotropy map. This is most simply done working in the Fourier representation: we draw the spherical harmonic coefficients $a_{\ell m}$ from uncorrelated Gaussians with zero mean and variance given by C_ℓ. Then we compute the spherical harmonic transform to obtain

$$T(\hat{n}) = \sum_{\ell=2}^{\ell=\infty} \sum_{m=-\ell}^{m=\ell} T_{\ell m} Y_{\ell m}(\hat{n}), \qquad (1)$$

where $Y_{\ell m}(\hat{n})$ are the spherical harmonics, an orthonormal and complete basis for functions on the sphere. The unit vector $\hat{n}$ points in a direction on the sky.

Unfortunately, the $T_{\ell m}$ or $T(\hat{n})$ cannot be observed directly. We are tied to our location in the Galaxy. There are various sources of copious amounts of microwave radiation in the Galaxy, such as dust and electrons from winds which are accelerated in the Galactic magnetic field. These foregrounds add to the CMB signal to make our sky.

This sky is then observed in various frequency bands (indicated by the red, green and blue arrows) by a CMB instrument (the figure shows an artist's conception of the WMAP satellite [33]). This instrument itself has a complicated transfer function: since weight and size constraints force satellite optics to be built

[3] http://www.rssd.esa.int/index.php?project=PLANCK

compactly the optics are not free from distortion. Microwaves have macroscopic wavelengths and therefore diffract around the edges of the instrument. This leads to sidelobes in the beam maps. The instrument scans the sky in a certain pattern (the "scan strategy") and internal instrument systematics are added to the scanned signal. The microwave detectors (either radiometers or bolometers) add noise to generate the time ordered data (TOD).

There are different levels of detail of the CMB data. To give an idea for the orders of magnitude involved, let us go through the sizes of the various objects in Figure 1 for the Planck experiment.

- The complete TOD for Planck will take up of order 1 Terabyte ($=10^{12}$ bytes) of storage (without counting house-keeping and pointing data).
- Each of the 100 detectors (channels) results in a map which is of order 10-100 Megabytes.
- The channels are grouped into of order 10 frequency bands.
- The combined maps at these different frequency bands will be combined into maps of the physical components (such as dust, synchrotron, CMB).
- The CMB power spectrum has a few thousand coefficients C_ℓ.
- These power spectrum coefficients are a function of $10-20$ cosmological parameters.

Note that there is a trade-off between the level of compression and what assumptions are implemented in the data analysis. Note also, that except for the raw data each of these data products by themselves mean little unless some means of assessing their statistical uncertainty is provided.

We immediately run into practical problems. For example, if we would like to specify a noise covariance matrix for each combined map at each frequency we would have to specify 10 times $\sim (10^6)^2/2$ elements of a matrix. The necessary storage space of order 10,000 Gigabyte basically precludes practical public distribution of the data. Other ways of specifying the uncertainty must be found.

Before we discuss in detail how cosmological information is encoded in the data, let us comment about an aspect of CMB statistics that is challenging even before there is any data in hand. This aspect is *experimental design*. While our experimental colleagues are putting a great deal of valuable thought into their instruments and observational strategies, it is currently still done in an informal way, by ingenuity rather than by formal method. This is partially so because it is hard to define optimality criteria that everyone would agree with and partially because of the immense effort involved to actually carry out the optimization. Evaluating any one proposed design requires many simulations of the full process from observation to analysis—the very process that presents the challenges I am discussing in this talk. Doing this repeatedly to search the space of design parameters for an optimal solution would be an immense task. Asymptotic techniques have been implemented on the basis of the Fisher matrix minimum variance predictions, but it is important to keep in mind that these are lower bounds on the expected variance, assuming a unimodal, Gaussian likelihood shape.

4. Inference From the Data

A stochastic model of the data and the information contained in it can be summarized in the following equations. For simplicity we will limit the discussion to a single channel. The TOD, d say, is modeled as the result of the action of a linear operator A which encodes the optics and scanning strategy of the instrument, on the sky, made up of signal s and foregrounds f:

$$d = A(s + f) + n. \tag{2}$$

Our assumptions is that s is an isotropic Gaussian random field with zero mean and power spectrum C_ℓ. We will take the noise correlations $N = \langle nn^T \rangle$ and A as given—though one of the statistical challenges of CMB analysis is to relax this assumption to some degree. We will not discuss this further here.

The task is then to extract as much information as possible about the cosmological parameters Θ from the TOD. To set up this inverse problem we write down Bayes' theorem

$$P(s, C_\ell, f, \Theta|d)P(d) = P(d|f, s, C_\ell, \Theta)P(f, s, C_\ell, \Theta). \tag{3}$$

In the Bayesian context, solving the inverse problem means exploring and summarizing the posterior density $P(s, C_\ell, f, \Theta|d)$. On the right hand side we can use the conditional independence of the C_ℓ and the data given s, and the plausible independence of f and s to simplify

$$P(f, s, C_\ell, \Theta) = P(f)P(s|C_\ell)P(C_\ell|\Theta)P(\Theta). \tag{4}$$

Traditionally, inference is performed in a linear sequence of steps which are concatenated into a pipeline. At each individual step a likelihood is written for the data in its current representation (e.g. TOD, maps, C_ℓ) in terms of the parameters describing the next stage of compression. Due to the complexity of evaluating the likelihoods, often the likelihood approach is abandoned and approximate, suboptimal but unbiased estimators are constructed.

Within the above framework we can understand each of the steps as a limit of Eq. 3. For example the "map-making" step is signal estimation with

$$P(f, s, C_\ell, \Theta) = P(f)P(s|C_\ell)P(C_\ell|\Theta)P(\Theta) = const. \quad (5)$$

Similarly, "Power spectrum estimation" summarizes the marginal posterior $P(C_\ell|d)$, where the d is taken in compressed form as the estimate of the CMB component. The summary proceeds through a maximum likelihood estimate (MLE) and the evaluation of the curvature of $P(C_\ell|d)$ at the MLE. In fact in most practical cases a non-optimal estimator is used due to the computational complexity of evaluating $P(C_\ell|d)$ or its derivatives and hence of computing the MLE.

Finally, "parameter estimation" signifies the step from the MLE of C_ℓ to a density $P(\Theta|d)$. The posterior $P(\Theta|d)$ is usually summarized in terms of the marginalized means and variances of the parameters or in terms of two-dimensional projections or marginalizations through the n-dimensional parameter space.

Recently, we developed MAGIC, a method that shows promise for global inference from the joint posterior [15]. The method, based on iterative sampling from the joint posterior, exploits the full dependence structure of the different science products from a CMB mission. For example it is useful for the separation of signal and foregrounds if the covariance of the signal is known. The information obtained on C_ℓ can be fed back into the CMB map estimate, which would in turn lead to a better estimate of C_ℓ, etc. But since I discussed this technique in my contributed talk [14] I will not go into details here.

I will now discuss a selection of the challenges we face in the "map-making," "power spectrum estimation" and "parameter estimation" steps on the way from the TOD to Θ.

5. Challenges in CMB Analysis

5.1. Map-making Challenges

Pixelizations of the sphere. A very basic challenge in CMB analysis is the fact that the CMB is a random field on the *sphere*. Convenient numerical techniques for storing and manipulating functions on the sphere needed to be developed. For CMB analysis in particular, fast methods for spherical harmonic transforms had to be developed and implemented. Point sets were needed that had good generalized quadrature properties such that discrete numerical approximations to Eq. 1 and its inverse

$$a_{\ell m} = \int d^2\hat{n} Y_{\ell m}(\hat{n}) T(\hat{n}) \quad (6)$$

generate accurate results. For local operations such as nearest neighbor searches and multi-resolution work, pixelizations of the sphere that allow hierarchical refinement are useful.

Various pixelizations of the sphere have been proposed [23, 24, 25]. Of these, HEALPix features pixels with exactly equal areas throughout, approximate equidistribution of pixel centers over the sphere, simple analytical equations for the pixel boundaries, fast spherical harmonic transforms and very favorable quadrature properties. Due to these features, HEALPix has developed into the standard pixelization for astrophysical all sky maps.

Beam Deconvolution Map-making. It is easy to show that the MLE $\hat{m}$ for the map $m = s + f$ is the result of maximizing Eq. 3 using Eq. 5. This results in

$$m \equiv (A^T N^{-1} A)^{-1} A^T N^{-1} d \quad (7)$$

with the associated noise covariance matrix $C_N =< mm^T >= (A^T N^{-1} A)^{-1}$.

These equations are generally valid whatever the forms of A and N. However, the non-trivial structure in the observation matrix A which is a consequence of the unavoidable imperfections in satellite optics makes solving for the map challenging. It is important to deconvolve the beam functions, since the beam distortions and side lobes lead to spurious shadow images of bright foregrounds. These images are spread throughout the map in a complicated way that depends on the scanning strategy. In addition, not accounting for beam imperfections results in distorting the signal itself. Both of these effects can bias not just the map but also the result of the covariance estimation.

For polarization map-making this deconvolution becomes even more important, since the polarization signal is weak and beam asymmetries can introduce spurious polarization signals into the data. Further, if bright foregrounds are significantly polarized, they may induce a significant polarization through their shadow images if beam convolution effects are neglected in map-making. Making high-quality maps of the polarization of the CMB anisotropy is the next frontier in CMB map-making.

To simulate the effects of realistic beams we have to have a general convolution technique for a beam map with a sky map along the scan path. If implemented as discrete sums over two pixelized maps, rotated to all possible relative orientations, such a technique requires of order $n_p^{2.5}$ operations, where n_p is the number of pixels in the maps. By doing the convolution in spherical harmonic space it was shown in [21] that this can be reduced to n_p^2 in the general case and to $n_p^{3/2}$ in interesting limit cases. This fast convolution method was generalized to polarization in [22].

Component separation. Once a map of the sky is made from each channel, these maps can be compressed losslessly into maps at each frequency. A great

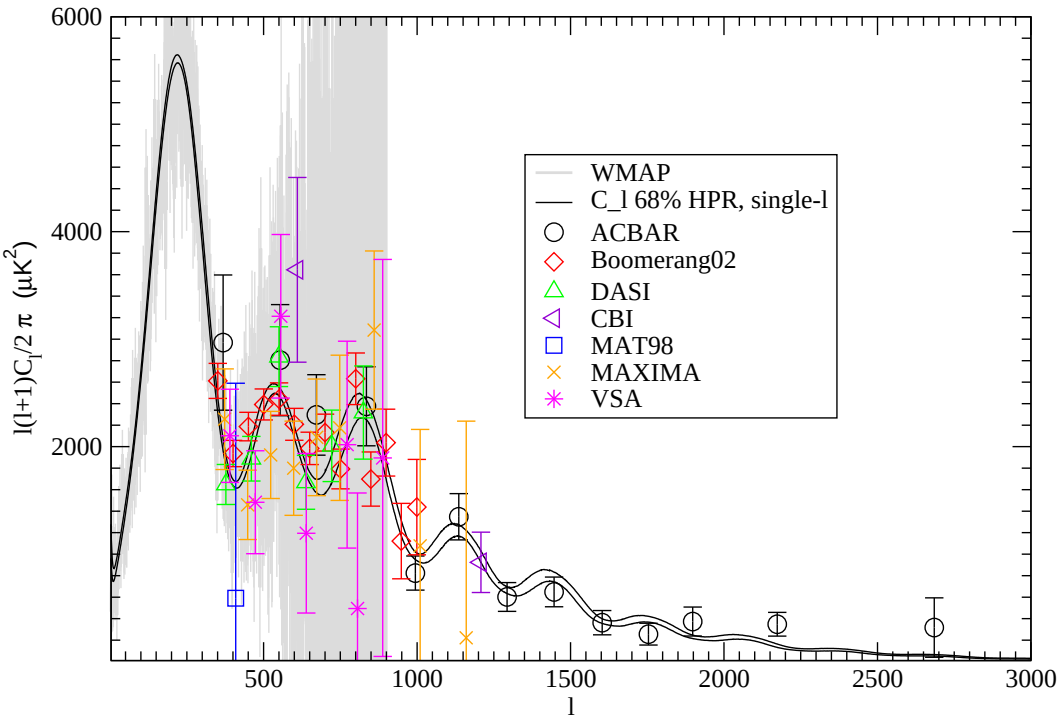

Figure 2: Our compilation of all recent CMB power spectrum (C_ℓ) data (points), including the WMAP data (gray band). Pre-WMAP data which provide redundant information about the power spectrum at low ℓ are omitted. Also shown (as two solid lines) is the 68% constraint on the power spectrum after implementing a prior which restricts the range of theories to a 10-parameter space of adiabatic inflationary theories (from [18]).

deal of work in the field has gone into devising methods for then obtaining an estimate of the CMB sky from these foreground contaminated maps at each frequency band, both for temperature and for polarization. We can either choose to model the foregrounds physically (e.g. [30, 31]) or we can attempt "blind separation," by defining an algorithm for automatic detection of different components in the maps, e.g. based on the statistical independence of these components (e.g. [29]).

Lensing. The CMB has traveled past cosmological mass concentrations which perturb the photon geodesics. This distortion contains valuable and complementary information about cosmological parameters. At the same time it mixes the polarization modes in the CMB data, contaminating the primary signal for the detection of the primordial gravitational wave background. Methods need to be devised that can measure this distortion and extract the information contained in it. It is intriguing that [32] find that exact techniques have significant advantages over approximate, quadratic estimators for the reconstruction of B polarization maps from lensed CMB.

5.2. Power Spectrum Estimation

The Computational Problem. For perfect (all-sky, pure signal, no noise) data, power spectrum estimation would be easy. Just compute the spherical harmonic transform, Eq. 6. Then the estimator

$$\hat{C}_\ell = \frac{\sum_m |a_\ell|^2}{2\ell+1} \tag{8}$$

is the MLE for C_ℓ. It is also easy to evaluate and explore the perfect data posterior to quantify the uncertainty in the estimates. However, in the general case we would like to evaluate $P(S(C_\ell)|d,N)$ which is the result of integrating out ("marginalizing over") s in the joint posterior. We obtain[4]

$$P(S(C_\ell)|d,N) = G(m, S(C_\ell)+C_N). \tag{10}$$

Here S is the signal covariance matrix, parameterized by the C_ℓ. Since $S+C_N$ is not a sparse matrix and since the determinant in the Gaussian depends on S, evaluating $P(S(C_\ell)|d,N)$ as function of the C_ℓ costs of order n_p^3 operations.

For Planck, $n_p \sim 10^7$ so if the constant factor in the scaling law was 1 (an unrealistic underestimate)

[4] We use $G(x,X)$ as a shorthand for the multivariate Gaussian density

$$G(x,X) = \frac{1}{\sqrt{|2\pi X|}} \exp\left(-\frac{1}{2}x^T X^{-1} x\right). \tag{9}$$

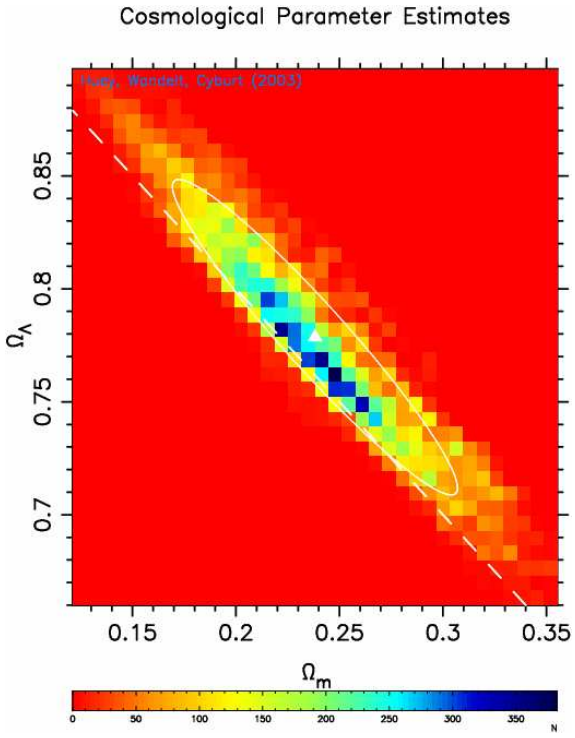

Figure 3: An example result of an online exploration of the marginal posterior of the dark matter density (Ω_m) and dark energy density (Ω_Λ) in our Universe. All recent CMB data, including the WMAP data, as well as the Hubble Space Telescope key project results and the Supernova cosmology project were included to obtain these constraints. Points in the figure are colored (from red to blue) according to how well the parameter combination at that point agrees with the data.

this would mean 10^{21} operations for one likelihood evaluation. For a 10 GFLOP CPU this means 1000s of CPU years of computation.

Various approaches have been suggested to to counter this challenge. They can be broadly divided into two classes: 1) specialized exact (maximum likelihood estimation) algorithms exploit advantageous symmetries in the observational strategy of a CMB mission [9, 10] to reduce the computational scaling from order n_p^3 to order n_p^2, and 2) approximate algorithms which filter unwanted properties of the data and simply compute the power spectra on the filtered and incomplete data, and then de-bias the results using Monte Carlo simulations after the fact[12]. This second class of algorithms is known as pseudo-C_ℓ algorithms[11] and has gained a great deal of popularity in recent analyses of CMB data, including the WMAP data [13].

Cosmic Variance. Aside from computational problems there is an interesting conceptual problem with CMB power spectrum analysis: the fact that we only have one sky.

This fact induces a fundamental limitation to how well we will be able to constrain cosmological parameter estimates from the CMB. The usual way of phrasing this limitation is in terms of cosmic variance.

Essentially, power spectrum estimation is variance estimation. For perfect data the solution that maximizes Eq. 10 is

$$\hat{C}_\ell = \frac{\sum_m |a_\ell|^2}{2\ell + 1}. \qquad (11)$$

So we are estimating the variance of the spherical harmonic coefficients. But, for example, for $\ell = 2$ there are only 5 such coefficients and a variance estimate from 5 numbers is statistically uncertain. So "cosmic variance" is expression of the fundamental limit to the precision of any measurement of the C_ℓ caused by the fact that the sphere is a bounded space and the fact that causality will not allow us to observe independent patches of the universe.

This fundamental limit to our knowledge provides a powerful motivation to do the best possible job in analyzing cosmological data.

5.3. Parameter Estimation

Techniques. The problem of parameter estimation is challenging since we need to explore the posterior density $P(\Theta|d)$ which varies over $\sim 10 - 20$ dimensions. Various techniques have been used to do this. For smaller number of dimensions (up to about 5) gridding techniques have worked well. However, the current state of the art is to use Markov Chain Monte Carlo methods [20] such as the Metropolis Hastings algorithm to sample from $P(\Theta|d)$ and to then base inferences on summaries of the posterior density computed from the sampled representation.

The question how to implement physical priors and constraints was one theme that was discussed at this conference. Within the Bayesian framework there exists a unique prescription for applying physical constraints through the specification of informative priors. The first example of a CMB parameter estimation which can be explored interactively online is the Cosmic Concordance Project [17]. This compiles data from several recent CMB observations and combines them with the user's choice of other, non-CMB experiments and physical priors. The result is displayed as the 2D marginal posterior density for any 2 parameters chosen by the user. An example is displayed in figure 3. I invite you to have a look at our prototype implementation at http://galadriel.astro.uiuc.edu/ccp.

A first scientific result from this project was a measurement of the fraction of ^{4}He in the Universe, both from CMB data alone and in the combination of standard Big Bang Nucleosynthesis with the CMB data. Since standard BBN links the primordial ^{4}He abundance to the baryon to photon ratio which is determined exquisitely well by CMB data, we obtain the

most precise measurement of the primordial ^{4}He abundance to date [18].

Several open questions remain to be addressed. What number n of parameters is needed to fit the data? Which selection of parameters from the full set should we use in the analysis? These questions are currently handled on the basis of personal preference of the authors or computational convenience—a statistically motivated procedure for letting the data selecting certain parameters has not yet been implemented.

Operationally, it is very computationally expensive to sample from a posterior with 10 parameters or more, because each likelihood evaluation requires running a Boltzmann code which computes $C(\Theta)_\ell$. In many dimensions the Metropolis sampler produces correlated samples (regardless of whether the target density is correlated or not—the correlations come from the sampler taking small steps through the many dimensional space). New numerical techniques for evaluating and sampling from the likelihood in high-dimensional spaces are needed.

Beating cosmic variance. In Figure 2 we show another application of parameter estimation techniques which is relevant to several themes we touched on in this talk and at this conference. The figure shows our compilation of CMB data and the mean and $\pm 68\%$ errors on the range of power spectra which are contained in our 10 parameter fits. In other words, we have implemented the optimal non-linear filter for the C_ℓ if the Universe is really described by our 10 parameter model. The physical prior has reduced the cosmic variance error bars far below the limit set by Eq. 11. Since we are using the physical prior that the CMB power spectrum is the result of plasma oscillations in the primordial photon baryon fluid the resulting smoothness of the power spectrum is used in constructing the estimate.

6. Testing the Assumptions: Challenges of Non-canonical CMB Analysis

In addition to the canonical analyses outlined above it is fundamentally important to test the assumptions on which these analyses rest. I will very briefly mention two of them here.

Is the CMB signal really Gaussian? Is it isotropic? These questions touch simultaneously on the issue of hypothesis testing, as well as model selection. What is the evidence in the data for the assumptions of isotropy and Gaussianity on which canonical analyses of the CMB are built? Can a goodness of fit criterion be defined which allows assessing whether the standard model of cosmology is a complete description of the CMB data?

The idea of testing the goodness of fit can be extended to cross tests of CMB data with other cosmological data. Ultimately the goal of the Cosmic Concordance Project is to allow users to select various data sets and explore interactively whether the parameter constraints from various data sets are compatible with each other. If this agreement is established the the constraints can be combined to generate yet stronger constraints. If disagreement is found this is motivation for observational groups to collect more data or for theoretical groups to work out new mechanisms that can reconcile the discrepant observations.

Testing for statistical isotropy in the CMB is a well-defined operation, since it is easy to specify alternative models. It requires checking whether a model of the correlations in the fluctuations in terms of rotationally invariant quantities (such as the power spectrum) is better or worse than a model that contains quantities that do not transform as scalars under SO(3). A frequentist test statistic has been suggested in [34], but a Bayesian treatment has not yet been attempted. The detection of a significant deviation from statistical isotropy would be a very important result, since isotropy is a fundamental prediction of inflation.

Testing for non-Gaussianity (NG) is similarly important, but much less well-defined. In the absence of physical models for NG a Bayesian treatment is not possible. The standard approach is to define some NG statistic, e.g. the skewness of the one-point function of the CMB fluctuations. Then this statistic is applied to the data and to a sample of Gaussian Monte Carlo samples (which are usually constrained to match the two-point statistics of the observed data). This Monte Carlo sample represents the null-hypothesis. A way is defined to assess discrepancy of the data and the Monte Carlo sample and if the discrepancy is statistically significant a detection is claimed.

Even though this is a straightforward frequentist procedure there is a great deal of arbitrariness in choosing the test statistic. Usually the choice is made based on a vague notion of genericity, in the sense that, for example, the skewness of the data is probably a more generic NG statistic than the temperature in pixel number 2,437,549, say. However, it is clearly not trivial to define a proper measure on the space of all possible statistics. The arbitrariness in the selection of the statistic (topological features of the map, n-point functions in pixel and in spherical harmonic space, wavelets, excursion sets, etc.) makes the test results difficult to interpret. For example, is clearly true that there will always be *some* statistic that gives an n-sigma result for any n!

To circumvent this arbitrariness, we could test for the NG predicted in certain variants of inflationary models ([26, 27, 28] and references therein). These physical models allow a Bayesian analysis in principle, and it would provide the best possible constraints on the presence of NG, but implementing Bayesian NG inference these models is computationally tedious.

7. Conclusion

In this talk I reviewed some of the more severe statistical and computational challenges of CMB analysis. CMB data are fundamentally important to cosmology, the problems I outlined are intellectually rich, and the knowledge that we need to make the most of the one CMB sky we can observe and analyze, present a powerful motivation to solve the CMB analysis problem.

Acknowledgments

I thank the organizers of PHYSTAT2003 for making this conference so enjoyable and stimulating. This work has been partially supported by the National Computational Science Alliance under grant number AST020003N and by the University of Illinois at Urbana-Champaign through an NCSA Faculty Fellowship.

References

[1] Mason, B.S., *et al.*, Ap. J. 591, 540 (2003)
[2] Seljak, U., and Zaldarriaga, M., Ap.J. 469, 437 (1996)
[3] Lewis,A., Ap. J. 538, 473 (2000)
[4] Guth,A., Phys. Rev. D. 23, 347 (1981)
[5] Linde, A., Phys. Lett. 108B, 389 (1982)
[6] Phys. Rev. Lett 48, 1220 (1982)
[7] C. L. Bennett *et al.*, Ap. J. 464, L1 (1996)
[8] J. R. Bond, A. H. Jaffe, and L. Knox, Physical Review D 57, 2117 (1998)
[9] S. P. Oh, D. N. Spergel, G. Hinshaw, Ap. J. 510, 551 (1999)
[10] Wandelt, B. D. and Hansen, F., Phys.Rev. D67, 023001 (2003)
[11] B. D. Wandelt, E. Hivon, and K. M. Górski, Phys. Rev. D 64, 083003 (2001)
[12] E. Hivon et al., Ap. J. 567, 2 (2002)
[13] G. Hinshaw, et. al., ApJS 148, 135 (2003)
[14] Wandelt, B. D., in these proceedings.
[15] Wandelt, B.D., Larson, D., and Lakshminarayanan, A., astro-ph/0310080, PRD in press.
[16] C. L. Bennett *et al.*, Ap.J.Suppl. 148, 1 (2003)
[17] Huey, G., Wandelt, B. D., and Cyburt, R., in preparation.
[18] Huey, G., Cyburt, R., and Wandelt, B. D., astro-ph/0307080, PRD in press.
[19] A. D. Milleret al. 1999 Astrophys.J. 524 (1999) L1-L4; N. W. Halverson et al. 2001, astro-ph/0104489; S. Hanany et al. Astrophys.J. 545 (2000) L5; Keith Grainge et al. Mon. Not. R. Astron. Soc. 000, 15 (2002); Kuo, C. L. et al. 2002, ApJ, astro-ph/0212289; J. E. Ruhl et al. 2002, astro-ph/0212229; S., Padin et. al., ApJ 549, L1, (2001); for a comprehensive list see http://lambda.gsfc.nasa.gov
[20] Christensen, N., *et al.*, Class. Quantum Grav. 18, 2677 (2001)
[21] Wandelt, B.D. and Gorski, K., Physical Review D63, 123002 (2001).
[22] Challinor, A. *et al* Physical Review D62, 123002 (2000).
[23] Górski, K.M., Hivon, E., and Wandelt, B.D., in "Evolution of Large Scale Structure from Recombination to Garching," Banday, A. *et al* (eds.), Garching, Germany (2000). See also the HEALPix homepage (http://www.eso.org/science/healpix).
[24] Crittenden, R.G. and Turok, N.G., astro-ph/9806374 (1998)
[25] Doroshkevich, A., *et al.* astro-ph/0305537 (2003)
[26] E. Komatsu, et.al., ApJS 148, 119 (2003)
[27]
[28] E. Komatsu, D. Spergel, and B. D. Wandelt, astro-ph/0305189 submitted to ApJ (2003)
[29] D. Maino, et al., MNRAS 334, 53 (2002)
[30] Stolyarov, V, et al., MNRAS 336, 97 (2002)
[31] Bennett, C. L., et al., ApJS 148, 97 (2003)
[32] Hirata, C. M. and Seljak, U., PRD 68, 083002 (2003)
[33] http://map.gsfc.nasa.gov/
[34] A. Hajian and T. Souradeep, ApJ 597, L5 (2003)